#3

JP99262

Express Mail # EK930786446US

JC892 U.S. PTO
09/740780
12/19/00

证　明

本证明之附件是向本局提交的下列专利申请副本

申　请　日：　1999 12 24

申　请　号：　99 1 26567.X

申请类别：　发明专利

发明创造名称：　词义消歧的机器翻译方法和系统

申　请　人：　国际商业机器公司

发明人或设计人：胡岗；蒋建民；裘照明；唐道南；杨力平

中华人民共和国
国家知识产权局局长　姜颖

2000 年 5 月 30 日

权 利 要 求 书

1. 一种词义消歧的机器翻译方法，包括步骤:

在对第一种语言的文本进行翻译时，如果根据上下文无法确定要翻译的字、词在第二种语言中的词义，则判断所述字、词或所述字、词所在的句子是否存在有关超链接的信息;

如果存在有关超链接的信息，则根据有关超链接的信息导出相关文本，基于所述相关文本将所述字、词翻译成第二种语言。

2. 根据权利要求1的词义消歧的机器翻译方法，其中所述第一种语言的文本是以HTML语言在万维网上公布的网页，而所述有关超链接的信息用于描述网页之间或同一网页各部份之间的联系。

3. 根据权利要求1的词义消歧的机器翻译方法，其中所述第一种语言的文本是PDF文件、Lotus Notes文件、Microsoft Word文件或Microsoft Windows help文件。

4. 根据权利要求1的词义消歧的机器翻译方法，其中基于相关文本将所述字、词翻译成第二种语言的步骤，进一步包括:

在所述相关文本中寻找包含所述字、词的母词组;

对每一母词组在第二种语言中的词义进行概率分析，选择合适的词义作为所述字、词的翻译结果。

5. 根据权利要求4的词义消歧的机器翻译方法，其中所述寻找母词组的步骤包括步骤:

对所述字、词所在句子进行语法分析，得出所述句子的语法结构;

在组合相关语法成份、检索翻译词库的基础上确定母词组;

由所有母词组在翻译词库中对应的相关词条组成临时词库用于后续的翻译步骤。

6. 根据权利要求4的词义消歧的机器翻译方法，其中所述概率分析包括同义词分析和固定搭配分析。

7. 一种词义消歧的机器翻译系统，用于将第一种语言的文本翻

译成第二种语言的文本，所述系统包括：词典、词义分析器、词法分析器、句法分析器、语法分析器、语义分析器以及翻译结果输出器，所述系统的特征在于还包括一个超链接信息处理器，用于在词义分析器根据上下文无法确定要翻译的字、词在第二种语言中的词义时，判断所述字、词或所述字、词所在的句子是否存在有关超链接的信息，如果存在有关超链接的信息，则根据有关超链接的信息取得相关文本，基于所述相关文本将所述字词或句子翻译成第二种语言。

8. 根据权利要求7的词义消歧的机器翻译系统，其特征在于所述第一种语言的文本是以HTML语言在万维网上公布的网页，而所述有关超链接的信息用于描述网页之间或同一网页各部份之间的联系。

9. 根据权利要求7的词义消歧的机器翻译系统，其中所述第一种语言的文本是PDF文件、Lotus Notes文件、Microsoft Word文件或Microsoft Help文件。

10. 根据权利要求7的词义消歧的机器翻译系统，其中超链接信息处理器通过以下处理将所述字、词翻译成第二种语言：

在所述相关文本中寻找包含所述字、词的母词组；

对每一母词组在第二种语言中的词义进行概率分析，选择合适的词义作为所述字、词的翻译结果。

11. 根据权利要求10的词义消歧的机器翻译系统，其中所述寻找母词组的处理包括：

对所述字、词所在句子进行语法分析，得出所述句子的语法结构；

在组合相关语法成份、检索翻译词库的基础上确定母词组；

由所有母词组在翻译词库中对应的相关词条组成临时词库用于后续的翻译。

12. 根据权利要求10的词义消歧的机器翻译系统，其中所述概率分析处理包括同义词处理和固定搭配处理。

8

说　　明　　书

词义消歧的机器翻译方法和系统

本发明一般涉及机器翻译技术，具体地说涉及基于超链接信息对词义消歧的机器翻译方法和系统。

机器翻译是利用计算机使一种文字或口语翻译成为另一种文字和口语的技术。即在语言学的关于语言形式和结构分析的理论基础上，依靠数学方法建立机器词典、机器语法，利用计算机巨大的存储容量和数据处理能力，在没有人工干预的情况下实现从一种语言到另一种语言（或另外多种语言）的自动翻译。机器翻译是一门涉及到语言学、计算机语言学和计算机科学等多门学科的边缘性应用学科。为了实现翻译功能，机器翻译系统必须具有词法分析、句法分析、语法分析、词典、成语词典、语义分析以及输出语言的能力．机器翻译系统包括转换型、知识型和语义型等几种类型，但实际应用时通常是综合运用这些类型的功能特性。

一般来说，目前实用的机器翻译系统可以实现语句级的机器翻译。对于一篇给定的文章，现有的系统可以通过静态分析上下文来选择一个合适的词义。

随着因特网的普及，仅仅通过静态分析上下文来选择合适词义的机器翻译系统已无法满足人们的需要．因为当用户通过Web浏览器来访问因特网上的站点时，他/她阅读的文章一般是以HTML（超文本标记语言）书写的网页．网页之间存在许多超链接。这样，当翻译系统试图翻译网页时，不应仅仅局限于静态分析上下文。当根据上文下无法确定或选择合适的词义时，可以通过动态分析超链接信息，来选择合适的词义。例如一个新的网页包含一些标题，这些标题具有超链接．其中一个题目为“Clinton wins senate support as Kosovo strikes near”．这里，我们假设源语言是英语，目标语言是中文．翻译系统很难将“strike”的中文词义确定为“袭击”．它可能

是“罢工”、“打”或“好球”。如果没有其它辅助信息可用，“strike”作为名词通常翻译成“罢工”。该标题所链接的详细内容是:

“President Clinton sought and won support from Congress for military action against Yugoslavia just hours after NATO ordered air strikes that could begin as early as Wednesday”。

在以上文本中，包含词组“air strike”。该词组只有一个词义“空袭”。在词组“air strike”中“strike”的词义是“袭击”。于是可以从词组“air strike”中确定标题中“strike”的词义。在大多数情况下，在一个主题中一个词只有一个词义。我们的发明就是基于这样的假设。

由此可以看出，尽管现有的机器翻译系统可以实现语句级的机器翻译，但在确定词义时，它们一般仅通过静态分析上下文来选择合适的词义，不能通过动态分析相关文本来提高翻译的准确性。对于因特网用户，这样的现有机器翻译系统是不可靠的。在以上例子中，用户感兴趣的是有关“袭击”这样的主题，而不是“罢工”，如果机器翻译系统将标题翻译成有关“罢工”这样一主题，用户就可能不再往下阅读了，从而错过有关“空袭”的详细内容。

为了解决以上问题，本发明提出一种基于超链接的对词义消歧的机器翻译方法和机器翻译系统。

根据本发明一个方面，提供一种对词义消歧的机器翻译方法，包括步骤:

在对第一种语言的文本进行翻译时，如果根据上下文无法确定要翻译的字、词在第二种语言中的词义，则判断所述字、词或所述字、词所在的句子是否存在有关超链接的信息;

如果存在有关超链接的信息，则根据有关超链接的信息导出相关文本，基于所述相关文本将所述字、词翻译成第二种语言。

根据本发明另一个方面，以上所述的第一种语言的文本是以HTML语言在万维网上公布的网页，而所述有关超链接的信息用于

描述网页之间或同一网页各部分之间的联系。

根据本发明的再一个方面，提供一种对词义消歧的机器翻译系统，用于将第一种语言的文本翻译成第二种语言的文本，所述系统包括：词典、词义分析器、词法分析器、句法分析器、语法分析器、语义分析器以及翻译结果输出器，所述系统的特征在于还包括一个超链接信息处理器，用于在词义分析器根据上下文无法确定要翻译的字、词在第二种语言中的词义时，判断所述字、词或所述字、词所在的句子是否存在有关超链接的信息，如果存在有关超链接的信息，则根据有关超链接的信息取得相关文本，基于所述相关文本将所述字、词或句子翻译成第二种语言。

由此可以看出，本发明的基本超链接对词义消歧的机器翻译方法和系统通过动态分析相关的超链接信息可以提高翻译的准确性。

通过以下结合附图对本发明优选实施例的详细描述，可以使本发明的优点、特征更加清楚。

图1为描述根据本发明一个优选实施例的对词义消歧的机器翻译方法的流程图；

图2为描述根据本发明一个优选实施例进行母词组收集过程的流程图；

图3为描述根据本发明一个优选实施例对词义进行概率分析的过程的流程图；和

图4为描述根据本发明一个优选实施例的对词义消歧的机器翻译系统的方框图。

在结合附图对本发明的优选实施例进行描述之前，首先对说明书中使用的技术术语进行说明。

· HTML:HTML的全称是“HyperText Markup Language”，中文名称是“超文本标记语言”，简单地讲它是所有的因特网站点共同的语言，所有网页都是以HTML格式的文件为基础，再加上一些其他语言工具（例如：JavaScript，VBScript，JavaApplet等）构成的。这些文件除了一些基本的文字外还包括一些标签（Tag），这些

标签由“<”和“>”符号以及一个字符串组成，如以下例子所示，而浏览器（如：Internet Explorer和Netscape）的功能是对这些标签进行解释显示出文字、图像、动画，播放出声音。

```
<HTML>
<HEAD>
<TITLE>网页标题</TITLE>
</HEAD>
<BODY BGCOLOR="#FFFFFF">
<P>这里是HTML文件的正文</P>
</BODY>
</HTML>
```

· URL：URL的全称是“Uniform Resource Locator”，中文的名称为“统一资源定位器”。简单地讲就是网络上一个站点、网页的完整地址。URL可以引导网络用户和应用程序获取来自不同网站、用不同协议传输的信息。

· HyperLink：超链接。正因为有了超链接，才能把Internet叫做互联网。网页上的超链接一般有三种：一种是与绝对网址（Absolute URL）的超链接，例如：从你的网站链接到IBM的主页：http//www.ibm.com；第二种称为相对网址（Relative URL）的超链接，例如将你的主页的一段文字或标题链接到同一网站的其它网页上去；还有一种称为同一网页的超链接，即书签。

· 母词组：一个词的母词组是包含该词的短语或固定搭配。例如“air strike”就是“strike”的母词组。

· 翻译词库：这是一个字、词/短语级的英/汉双向字典。

以下结合附图描述根据本发明的优选实施例。

如图1所示，根据本发明一个优选实施例的基于超链接对词义消歧的机器翻译方法包括：

步骤101：去标签。为了对第一种语言的本文进行机器翻译，一般在翻译之前应将所有标签去掉，即去掉为了在网上公布文本而人为加上的那些标记。将所有去掉的标签放入标签库中。一般对本文翻译之后，在回送翻译结果之前要重新加上标签，这样浏览器就可以对这些标签进行解释显示出第二种语言的文本。

步骤102：超链接检验。该过程用于检查一个句子或一个句子中的某些词是否具有超链接。对于以HTML语言公布的文本，超链接可以是当前网页的命名地址或用作网址的URL。要处理的URL应满足：

1. 网络协议必须是HTTP协议；

2. Content_Type必须是text/html.

例如一个具有超链接的句子可以是：

I love this <A HREF= "http://abc.xyz.net/game.html">game</A>

除了Web网页，本发明的机器翻译方法也适用于其它具有超链接信息（如带有链接或书签）的文本。例如Adobe Portable Document Format (PDF)文件，Lotus Notes文件，Microsoft Word文件，Microsoft Windows helps文件等。正如本领域技术人员公知的那样，只需改变检查超链接步骤，本发明的方法就可以适用其它类型的文件。

例如对于Microsoft Word RTF文本，超链接可以是文件中的书签（bookmarks）或URL。这样上述的例子用RTF格式可以写成：

I like this {\field {\fldinst HYPERLINK "http://abc.xyz.net/ game. html"} {/fldinst game}}.

步骤103：收集母词组。在基于超链接信息取得相关文本后，可以在相关文本中逐句收集母词组。具体过程稍后将结合附图2进行详细的说明。

值得注意的是，相关文本可以来自第一级超链接，也可以来自多级超链接。

步骤104：概率分析。该过程对每一词义进行概率分析。具体

分析过程稍后将结合图3进行详细说明。

步骤105: 选择词义。在稍后将对此过程进行详细说明。

步骤106: 加上标签并输出翻译结果。

由以上可以看出本发明的对词义消歧的机器翻译方法可以动态分析超链接信息，基于超链接信息获取相关文本，由相关文本中的母词组确定要翻译的字、词的词义，这样就提高了机器翻译的准确性。

以下结合图2介绍一下母词组的收集过程。如图2所示，对于相关文本中的每一句子（201），首先通过语法分析给出句子的语法结构（202）。对于前述的例子，子句“NATO ordered air strike”可以分析为

NATO	order	air strike
subject	verb	object

多字词“air strike”是一个语法成份（203）。在此情况下，“air strike”可以认为是“strike”的母词组。在翻译词库中，寻找对应的词条（204），对应词条为:

air strike <n(ac)<t(空袭)<o(次)<x(,袭击)

在该词条中，“n”表示“air strike”是一个名词。“ac”表示“action”，这是它的语义分类。“t(空袭)”表示将它翻译为“空袭”。“o（次）”表示在中文中其度量单位为“次”。最后，“x(,袭击)”说明“air strike”是“strike”的母词组，而“strike”的词义为“袭击”并且“air strike”不是“air”的母词组。

对于例子“He received his doctoral degree”进行语法分析（202），得出语法结构为:

he	receive	his	doctoral	degree
subject	verb	pron-->	adj-->	object

“his”和“doctoral”是词“degree”的两个修饰词。根据语法规则可以将这样的相关语法成份组合起来（203）。在翻译词库中选择下述词条（204）:

receive <v(Obj degree/)<t(获得*学位)<x(学位,ab,个)

在该词条中，“v”说明这是一个动词短语。“(Obj degree/)”表示动词“receive”需要“degree”作为其宾语。“/”表示词“degree”是该词条的头词。“t(获得*学位)”是该词条的翻译结果并且“*”可以由“degree”的修饰词来代替。“x(学位,ab,个)”表示头词“degree”的信息：其翻译结果（学位）、语义类别（抽象宾语）和度量单位（个）。于是“receive(v)degree(obj)”是“degree”的母词组。

通过处理相关文本可以获得所有的母词组。由具有以下词条的那些母词组构成临时词库（206）用于后续的翻译。

strike<n(ac)<t(袭击)<o(次)

degree<n(ab)<t(学位)<o(个)

以下参照图3描述一个词义的概率分析过程。如图3所示，词义的概率分析过程（300）分为同义词分析（301）和固定搭配分析（302）两个子过程。

1. 同义词分析:

对于每个待消歧的词义，在相关文本中寻找其同义词。如果找到一同义词，则增加了是该词义的可能性。例如，对于句子“China erupts in fary at NATO strike”中的词“strike”，相关文本中有一句子:

“Thousands of people protested at U.S. and other diplomatic missions in China Saturday in an officially sanctioned outpouring of fury at NATO's bombing of the Chinese embassy in Belgrade.”

对于中文的词义“袭击”，有下列同义词:

“侧击 冲击 出击 打 动武 发 发射 反攻 反击 反扑 伏击 攻 攻打 攻击 攻坚 合击 轰击 轰炸 还击 还击 回击 回击 火攻 击 夹攻 夹击 截击 进攻 进击 狙击 开 空袭 拦击 炮击 破击 奇袭 枪击 强攻 强占侵袭 射击 投弹 突击 围攻 围击 袭击 掩杀 掩袭 佯攻 邀击 夜袭 炸 主攻 助攻 追击 阻击”.

在相关文本中，词“bombing”的词义为“轰炸”，这是“袭击”的同义词。于是，“strike”的词义为“袭击”的可能性要比是其它词义的可能性大。为此对于相关文本中每个关键词都要搜索一遍同义词库。如果找到一同义词，则该词义的加权值就会增加。例如对于上述的例子，“strike”的初始加权值为：

词义	加权
罢工	40
袭击	33
打	21
好球	2

其中的加权值是每个词义在大的语料库中出现的频率。在经过同义词分析后，加权值变为：

词义	加权值
罢工	40
袭击	33 + 10
打	21
好球	2

2. 固定搭配分析：

对于源语言进行语法分析可以给出句子的局部句法关系。例如，待消歧的词是句子“He developed the film”中的“develop”。在翻译词库中，对于具有一个宾语的动词“develop”的词条有：

develop<v (obj1)<t(成长)

develop<v (obj1)<t(冲洗)

develop<v (obj1)<t(发展)

develop<v (obj1)<t(开发)

develop<v (obj1)<t(使obj1成长)

dev lop<v (obj1)<t(使obj1形成)

语法分析给出：“develop”和“film”具有动－宾句法关系。词“film”作为名词具有下述词义：

16

film<n (mm na)<t(薄层)<o()

film<n (mm)<t(薄膜)<o(片)

film<n (mm)<t(胶片)<o(张)

film<n (ab)<t(影片)<o(部)

film<n (ab)<t(影片业)<o()

在目标语言语料库中带有动－宾句法关系组合的频率为:

	薄层	薄膜	胶片	影片	影片业
成长	0	0	0	0	0
冲洗	1	2	7	0	0
发展	0	0	0	0	2
开发	0	1	2	0	0
使…成长	0	0	0	0	1
使…形成	0	0	0	0	0

其中，“冲洗－胶片”是具有最大可能性的搭配。

系统中使用的句法关系为: 动－宾、形容词－名词、主语－动词、副词－动词和名词－名词。

以下介绍一下如何通过匹配临时母词组词典、同义词词典、固定搭配词典以及翻译词典中的词条来选择合适的词义。一般是根据加权值来选择合适的词义。在以下例子中，由于在临时母词组词库中的词条的加权值比其它词库中的词条的加权值大，所以对于“strike”，我们选择母词组词库中的词义。

130	strike<n (ac)<t(袭击)<o(次)	母词组词库
43	strike<n (ac)<t(袭击)<o(次)	同义词词库
40	strike<n (ab)<t(罢工)<o(次)	翻译词库
33	strike<n (ac)<t(袭击)<o(次)	翻译词库
21	strike<n (ac)<t(打)<o()	翻译词库
2	strike<n (ab)<t(好球)<o(个)	翻译词库

以上以英语作为源语言（第一种语言）、中文作为目标语言（第二种语言）介绍了根据本发明一个具体实施例的基于超链接信

息对词义消歧的机器翻译方法。该方法已在一个称为“HomePage翻译器2.0”的在线网页翻译系统中实施了。该翻译系统的结构如图4所示。一般机器翻译系统包括：词典、词义分析器、词法分析器、句法分析器、语法分析器、语义分析器以及翻译结果输出器。根据本发明一个优选实施例的基于超链接信息对词义消歧的机器翻译系统还包括一个源文本分析器，即超链接信息处理器，用于在词义分析器根据上下文无法确定要翻译的字、词在第二种语言中的词义时，判断所述字、词或所述字、词所在的句子是否存在有关超链接的信息。如果存在超链接的信息，则根据有关超链接的信息取得相关文本，基于所述相关文本将所述字、词翻译成第二种语言。如图4所示，在词义分析器根据上下文无法确定词义时，源文本分析器参照超链接库确定是否存在超链接信息，如果存在的话，则取得相关文本，对相关文本逐句进行语法分析，由母词组收集器将收集到的母词组放入临时词库中用于后续的翻译工作。为了提高翻译的准确性，本发明的翻译系统还由同义词分析器对词义进行同义词分析、固定搭配分析器对词义进行固定搭配分析。同义词分析器和固定搭配分析器在进行概率分析时各自参照自己的词库，即同义词词库和固定搭配词库。在经过概率分析后，词义选择器从相应的词条中选择具有最高加权值的词义作为翻译结果。

由以上介绍可以看出，根据本发明的基于超链接的对词义消歧的机器翻译方法和系统可以提高机器翻译的准确性。

正如本领域一般技术人员所理解的，本发明能由各种不同于所描述的实施例的方式来实现，这些实施例只是为了说明的目的，并不是为了限制本发明，本发明保护范围应由权利要求书来限定。

图1

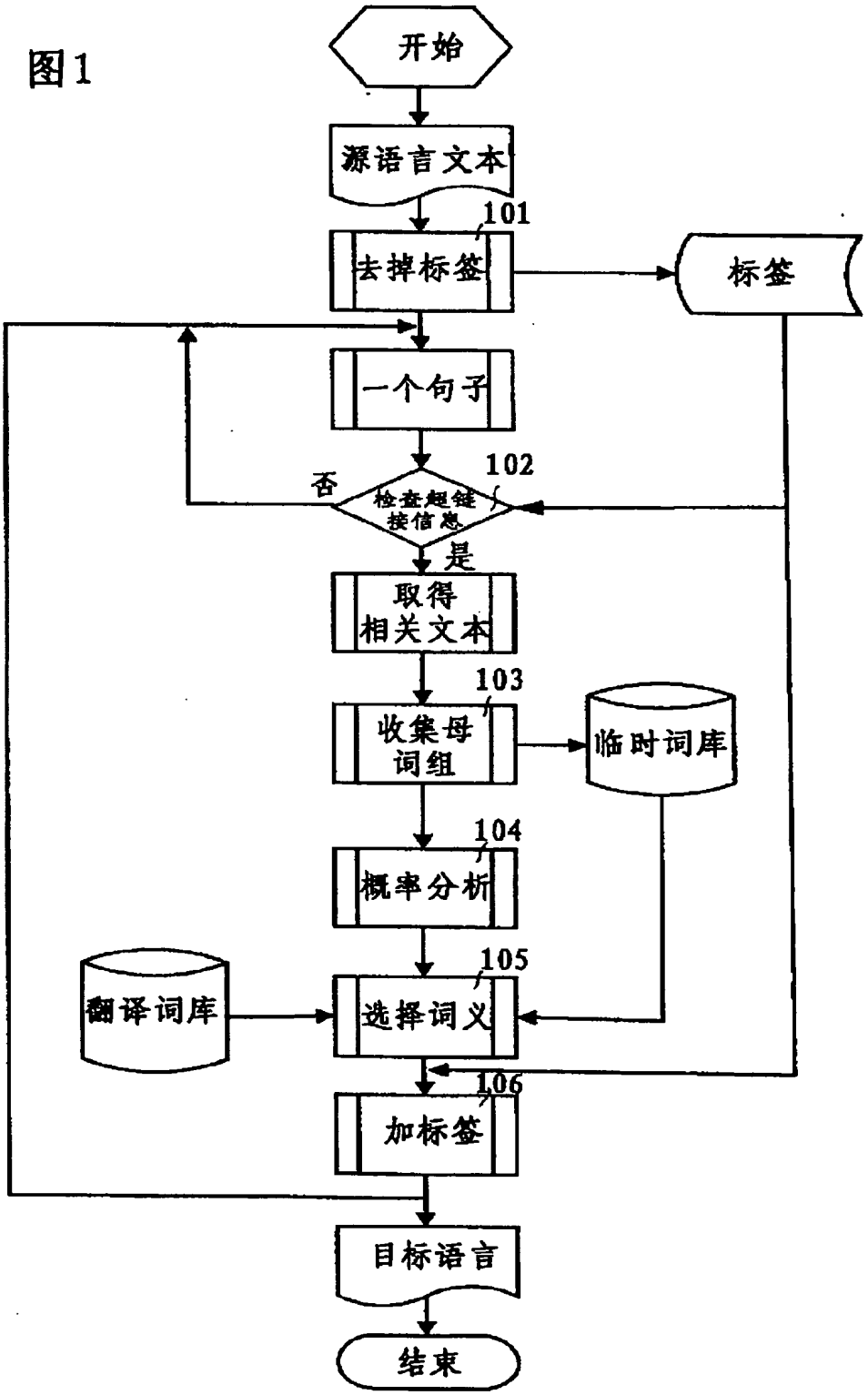
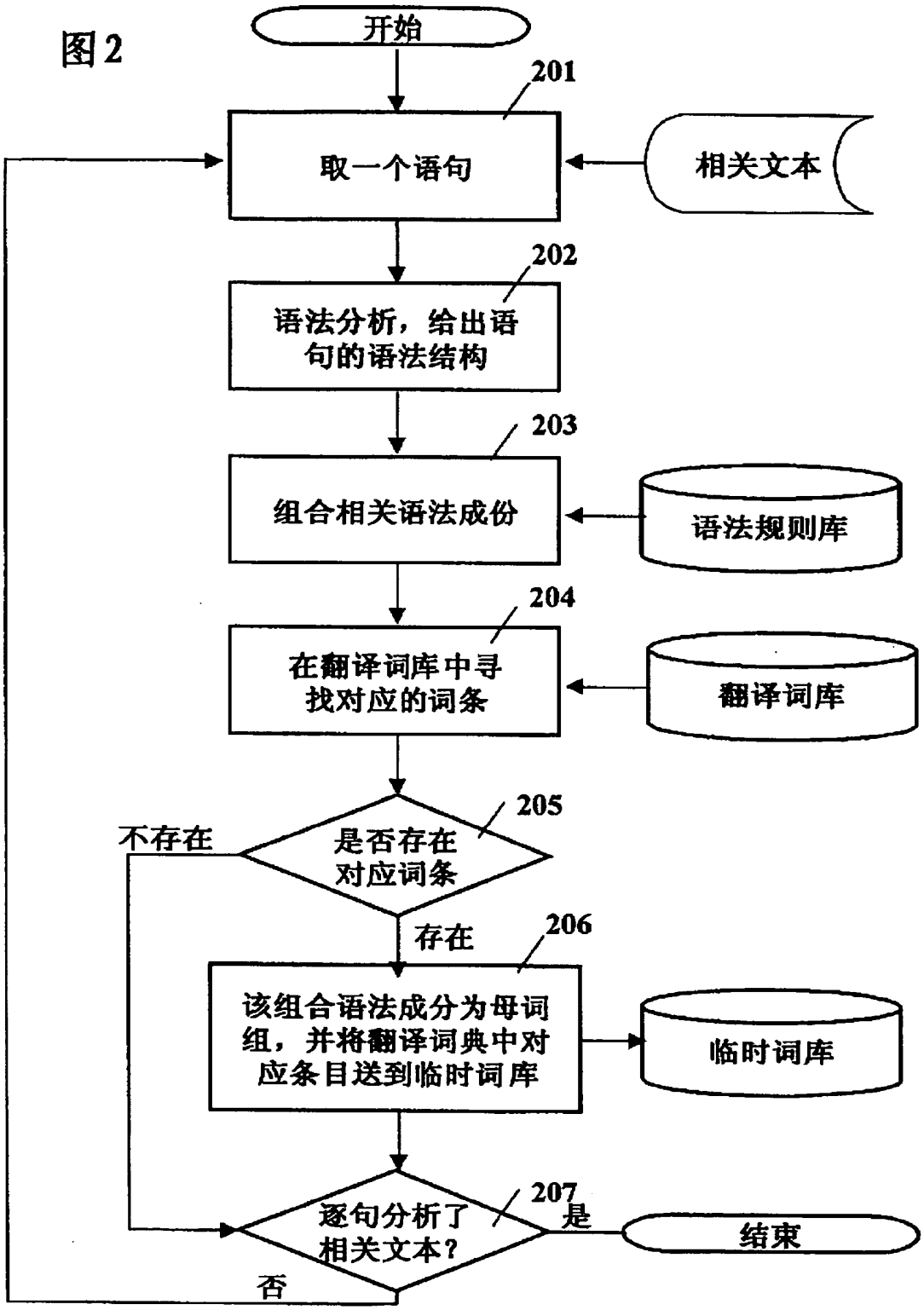

开始
源语言文本
101
去掉标签
标签
一个句子
102
否
检查超链接信息
是
取得相关文本
103
收集母词组
临时词库
104
概率分析
105
翻译词库
选择词义
106
加标签
目标语言
结束

图2


开始
201
取一个语句
相关文本
202
语法分析，给出语句的语法结构
203
组合相关语法成份
语法规则库
204
在翻译词库中寻找对应的词条
翻译词库
205
是否存在对应词条
不存在
存在
206
该组合语法成分为母词组，并将翻译词典中对应条目送到临时词库
临时词库
207
逐句分析了相关文本？
是
结束
否

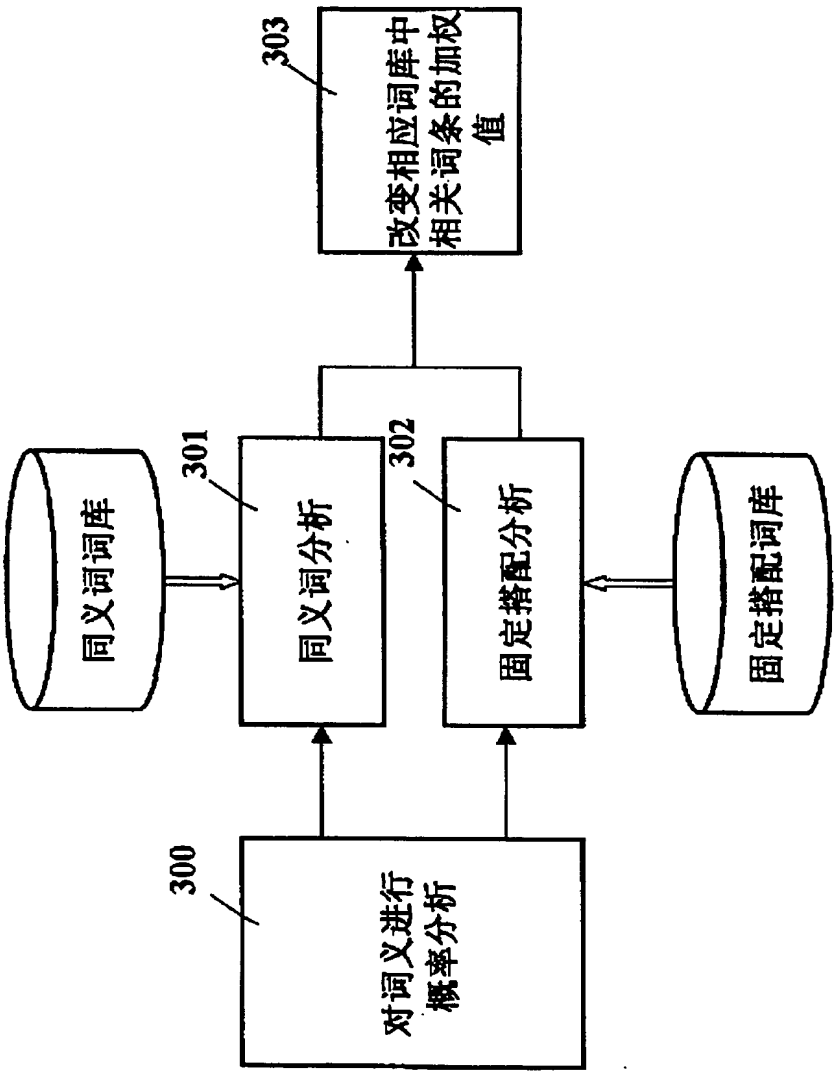
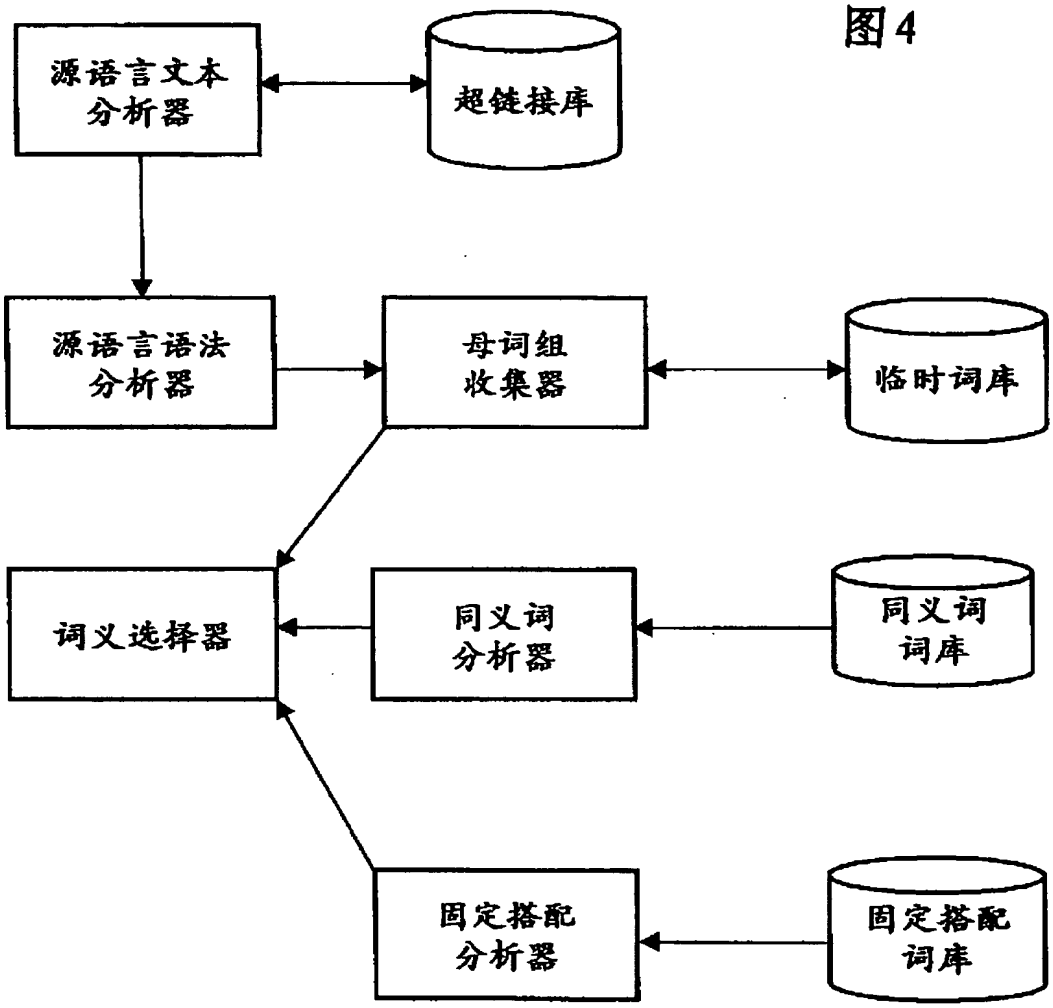

303
改变相应词库中相关词条的加权值
301
同义词分析
302
固定搭配分析
同义词词库
固定搭配词库
300
对词义进行概率分析

图3

图4


源语言文本
分析器
超链接库
源语言语法
分析器
母词组
收集器
临时词库
词义选择器
同义词
分析器
同义词
词库
固定搭配
分析器
固定搭配
词库